Gradient Inversion with Generative Image Prior

Year: 2021 | Citations: 182 | Authors: Jinwoo Jeon, Jaechang Kim, Kangwook Lee, Sewoong Oh, Jungseul Ok

---

Abstract

Federated Learning (FL) is a distributed learning framework, in which the local data never leaves clients devices to preserve privacy, and the server trains models on the data via accessing only the gradients of those local data. Without further privacy mechanisms such as differential privacy, this leaves the system vulnerable against an attacker who inverts those gradients to reveal clients sensitive data. However, a gradient is often insufficient to reconstruct the user data without any prior knowledge. By exploiting a generative model pretrained on the data distribution, we demonstrate that data privacy can be easily breached. Further, when such prior knowledge is unavailable, we investigate the possibility of learning the prior from a sequence of gradients seen in the process of FL training. We experimentally show that the prior in a form of generative model is learnable from iterative interactions in FL. Our findings strongly suggest that additional mechanisms are necessary to prevent privacy leakage in FL.